FACULTY OF SCIENCES

GHENT UNIVERSITY

SURVIVAL ANALYSIS PROJECT

A statistical Analysis on the Randomised International Stroke Trial Dataset to Investigate the Survival Time of Patients Treated With Aspirin

Academic Year 2020/2021

Group Members

Dereje Mengist Belete

&

Enyi Emmanuel Nfor

Table of Contents

1. Introduction

Acute stroke is known as the acute onset of focal neurological findings in a vascular territory as a result of underlying cerebrovascular disease. In this study, we analyse a dataset from a Randomised Trial of Aspirin and Subcutaneous Heparin among 19435 patients with Acute Ischaemic Stroke obtained from the International Stroke Trial (IST). The IST was a large randomized trial where patients after stroke onset were randomized to different combinations of Aspirin and Heparin treatments. The aim of our analysis is to examine the effect of aspirin on time to death, time to death due to specific causes, and to see the association between covariates such as sex, age systolic blood pressure, and stroke type with time to death.

2. Methodology

The seven studied variables were extracted from the entire dataset, with missing values removed from each covariate. It was presumed that missingness occurred at random and has no effects on treatment or our outcome. In addition, several variables were recoded and renamed for easier interpretation. The effect of aspirin on the outcome variable was assessed using a log rank test, and the finding was validated using Cox PH. Furthermore, the final Cox PH model was used to examine the relationship between each covariate and outcome variable. The study was carried out on the entire dataset without dividing it into training and test data. The PH assumption was tested using the Schoenfeld residuals test, and non-informative censoring was assumed throughout the study.

3. Result

3.1. Characteristics of Patients

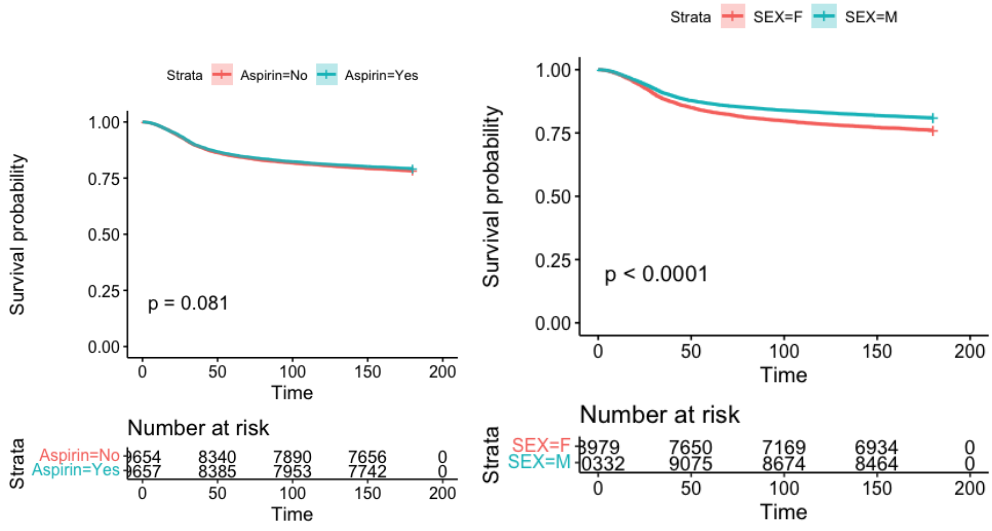


Fig 1 KM stratified by treatment (on the left) and KM stratified by gender (on the right)

This study involves 10332 male patients and 8979 female patients, with an average age of 73 years and an average Systolic blood pressure of 160 mm/hg. In six months, 3290 patients died, while 12159 were censored. Following the onset of a stroke, 9657 patients were randomly assigned to take aspirin, while 9654 were not. Patients with and without aspirin treatment seems to have the same survival curve, meaning that aspirin therapy had little effect on reducing mortality of patients with Acute Ischaemic Stroke. The KM curve reveals that the survival curves of the two genders vary, with males having a longer survival period than females. However, since it is difficult to tell if this difference is statistically significant (at the 5% level) based on this plot alone, a log rank test was used to test the difference. The null hypothesis states that there is no difference in survival curves between those who take aspirin and those that do not. The p-value for the log rank test is 0.08, indicating that there is no evidence of a difference in survival curves between stroke patients who take aspirin and those who do not. In addition, log rank was used to see whether there was a significant difference in gender groups. The null hypothesis is that the survival curves of the two genders are same. The p-value <0.0001 from log rank test implies there is a statistically significant difference in the survival curve between male and female patients. The proportional hazard was checked by plotting log (-log (S(t)) and the assumption seems to hold hence the log rank test power is optimal (Fig 2)

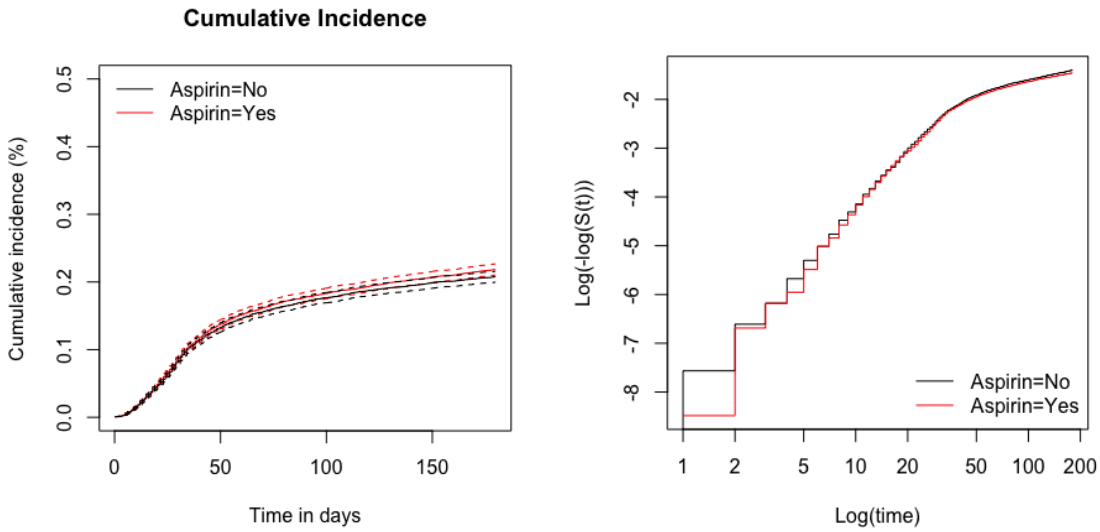


Fig 2: Cumulative incidence curve (on the left) and Complimentary log-log plot by treatment (on the right)

3.2. Effect of Aspirin Treatment on Time to Death

As seen in the cumulative incidence curve, there is no identifiable gap in mortality rates between patients who received aspirin treatment and those who do not. To be more specific,

we will use the curve to measure the hazard rate for each treatment arm and the hazard ratio for aspirin treatment. For example, the hazard at 100 days can be calculated as follows,

$\overline{\lambda}Asprin_Y = \frac{0.16}{100} * 180$ for 6 months per patients

$\overline{\lambda}Asprin_N = \frac{0.17}{100} * 180$ for 6 months per patient

The HR will be the ratio of the two-treatment arm which approximately is 0.95. To validate this, we have used a cox PH model and found HR=0.95 which shows the hazard of death is 5% lower for patients that takes aspirin than those that don't take aspirin at any given time, this difference however isn't significant at 5% with p-value 0.2 and 95% confidence interval (0.89, 1.02). Treatment effects were also compared using a cox model stratified by hospital. The baseline hazard for each hospital is assumed to be different if stratified by hospital, and the treatment effect is expected to be the same. The result showed HR of 0.95, with p-value =0.1 and 95% confidence interval of [0.8923, 1.01], which implies that survival curve for the two treatment groups isn't significantly different across each stratum, this result is in line with the log rank test and the cox model without stratifying by hospital.

To determine the effect of aspirin on specific cause of death i.e., death due to Pulmonary Embolism, the indicator variable for the specific cause of death was categorized into two events. One death due to Pulmonary Embolism as event one, and death from other causes as event two. The Gray test was used to test the following two hypotheses: first, the cumulative incidence for Death due to Pulmonary Embolism is the same for patients who receive aspirin treatment and those who do not receive aspirin treatment. Second, the cumulative incidence for Death due to other causes is the same for patients who take aspirin and others who do not take aspirin. The test indicated that there is no significant difference in death due to pulmonary embolism and death due to other causes between the treatment groups with p-value of 0.21and 0.14 respectively. However, as it depicted in the cumulative incidence plot more patients have died due to the other causes than due to pulmonary embolism (Fig 3).

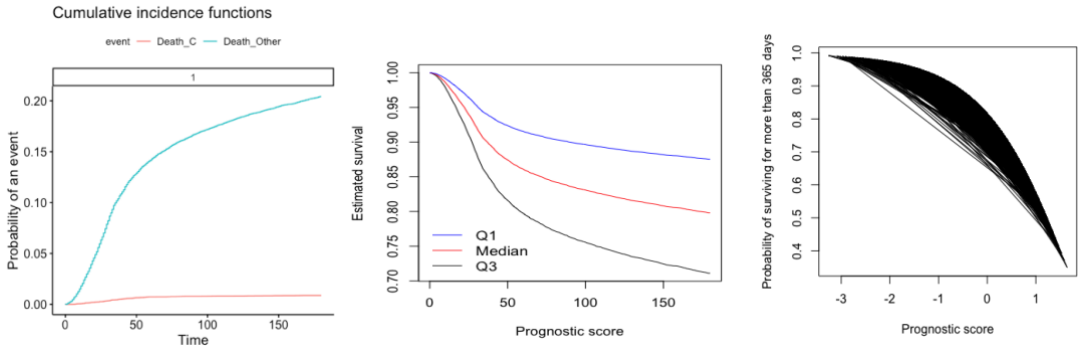


Fig 3 Cumulative Incidence Function (on the left), Estimated Survival (in the middle), Probability of surviving for more than 1 year (on the right)

Further, we have modelled the cause specific hazard for each event separately, for death due to pulmonary embolism (event 1) we found HR =0.82, p=0.208, and 95% CI [0.6, 1.1]. The HR is smaller than 1 implying the treatment is beneficial, however, the difference between the treatments to reduce the hazard isn't statistically significant. Similarly, the cause specific model for event 2, death due to other causes indicates that there is no statistical difference between treatment group with P= 0.129 and HR= 0.95 and 95% confidence interval [0.89, 1.01].

3.3. Advantages and Disadvantages of Adjusting for Covariates

Adjustment for other covariates can improves the efficiency of the analysis and hence produces stronger and more precise evidence (smaller p-values and narrower confidence intervals) of an effect. Further, adjusting for prognostic covariates in the analysis of time-to-event outcome when estimating a hazard ratio will also lead to an increase in power. On the other hand, covariate adjustment can lead to type I error rates (increased probability of a false positive) especially when there is a small sample size and time-to-event outcome, but our sample is large enough. Furthermore, when there is missing data on certain covariates and patients with missing values for the covariates are excluded from the analysis, the analysis becomes unsatisfactory as it will reduce the sample size, and therefore reduce power.

3.4. Model Building

The baseline group (reference group) in the analysis is defined as patients who are female, LACS stroke type, and patients who were not randomized to Aspirin treatment. We start building a Cox PH model by fitting separately for each of the predictors starting with age. The p-value for age is $p<0.0001$, with a hazard ratio HR = exp(coef) = 1.06, and 95% confidence interval [1.05, 1.06] indicating a strong significant relationship between the patients' age and increased risk of death. Therefore, patients with an additional year of age induce daily hazard of death by a factor of exp(beta) = 1.06, or 6%. This can be explained further as patients in the first quartile (65 years old) have more survival time than the median (73 years old) and third quartile (80 years old), as seen in Fig 3 of the survival curve.

The added effect of SBP can be explained by HR=0.99, 95% CI [0.995, 0.998] and $p< 0.0001$, indicating the hazard of death is 1% lower for patients per one unit increase in SBP, the significant p-value shows a strong relationship between the SBP and decreased risk of death, holding the age constant, a higher value of SBP is associated with a higher survival. The effect of age on hazard remains the same after adding SBP. To compare which of the two predictors,

in terms of magnitude, we could look the difference of their 3rd and 1stquartile range time their estimate since they are both continuous. Age =0.059 has larger estimate than systolic blood pressure (-0.0021). Furthermore, the AIC values of the two models for each covariate were compared, and the model with just age has an AIC of 78800.95, whereas the AIC model with SBP has an AIC of 78800.95, implying that the model with age is a better fit. Whereas, in terms of variability we can look the standard error and CI range for each covariate and age has higher standard error and more wider confidence interval which implies less precision and more variability compared to SBP.

After including other covariates, the estimate for SBP is found to be small, but the magnitude for continuous variables is measured as estimate multiplied by the difference between the first and third quartiles, as opposed to categorical variables, where the estimate in the result is the true effect. The interaction between treatment and stroke type isn't statistical significance with p value 0.38, therefore the interaction term was dropped from the model.

Table 1 Cox Final Model Estimates

	Estimates	HR (95%CI)	SE	P
Aspirin-Yes	-0.053	0.95 (0.89, 1.01)	0.031	0.088
Age	0.054	1.055 (1.05, 1.06)	0.002	0.0001
Sex-M	0.045	1.05 (0.98, 1.11)	0.032	0.157
SBP	-0.003	0.99 (0.99, 1.00)	0.001	0.0001
Stroke-PACS	0.74	2.10 (1.88, 2.35)	0.056	0.0001
Stroke-POCS	0.76	2.13 (1.86, 2.45)	0.071	0.0001
Stroke-TACS	1.52	4.60 (4.13, 5.13)	0.055	0.0001

The final model can be written as

$$\lambda(t;Z) = \lambda_0(t)exp(-0.053 * Z_1 + 0.054 * Z_2 + 0.045 * Z_3 - 0.003 * Z_4 + 0.74 * Z_5 + 0.76 * Z_6 + 1.52 * Z_7$$

Where Z_1..Z_7 refers to the covariates included in the model chronologically i.e. Z_1 is Aspirin yes and Z_7 is Stroke-TACS

The treatment/aspirin has HR= 0.95 with 95% CI [0.89, 1.01]. This can be interpreted as keeping the SBP and age covariates constant, for female patients with LACS stroke type and aspirin treatment the hazard will reduces by a factor of 0.95, or 5%, in comparison to male patients with other types of stroke and that weren't randomized into aspirin treatment.

However, the treatment contribution to the difference in the HR is small and association was not significant, p-value =0.08 at 5 %. The p-value for Sex is 0.157, HR = 1.05, with a 95% confidence interval [0.98, 1.11], this indicates that adjusted for other covariate male patients are associated with higher risk of death. Furthermore, we can infer from the p-value and the CI as it includes 1 that sex makes a smaller contribution to the difference in the HR. In addition, we can calculate the HR and CI of two patients with specific characteristics. For instance, the estimated hazard ratio for death for a patient aged 80 years versus a subject aged 65 years, HR= $exp(0.054 * 15)$= 2.24, indicates the hazard is more than two times higher for patients aged 80 than patient aged 65 years. The respective 95 % CI [2.12,2.25] can be computed as

$$[15\overline{\beta}age \pm 1.96 * 15 * \overline{SE}(\beta age)$$

Further, the estimated hazard ratio for death for a subject of age 80, with stroke type PACS randomized to aspirin, vs a subject of age 60, with stroke type TACS randomized to aspirin is 1.33, this show patients with older age and PACS type of stroke have more 33% death rate compared to younger patients with TACS type of stroke.The overall survival curve for the whole model (Appendix Fig 5) will reveal the average survival time for all covariates adjusted, but we can't see the magnitude and relationship of each covariate with survival time. The KM curve can be better in this respect to compare each covariate survival time separately.

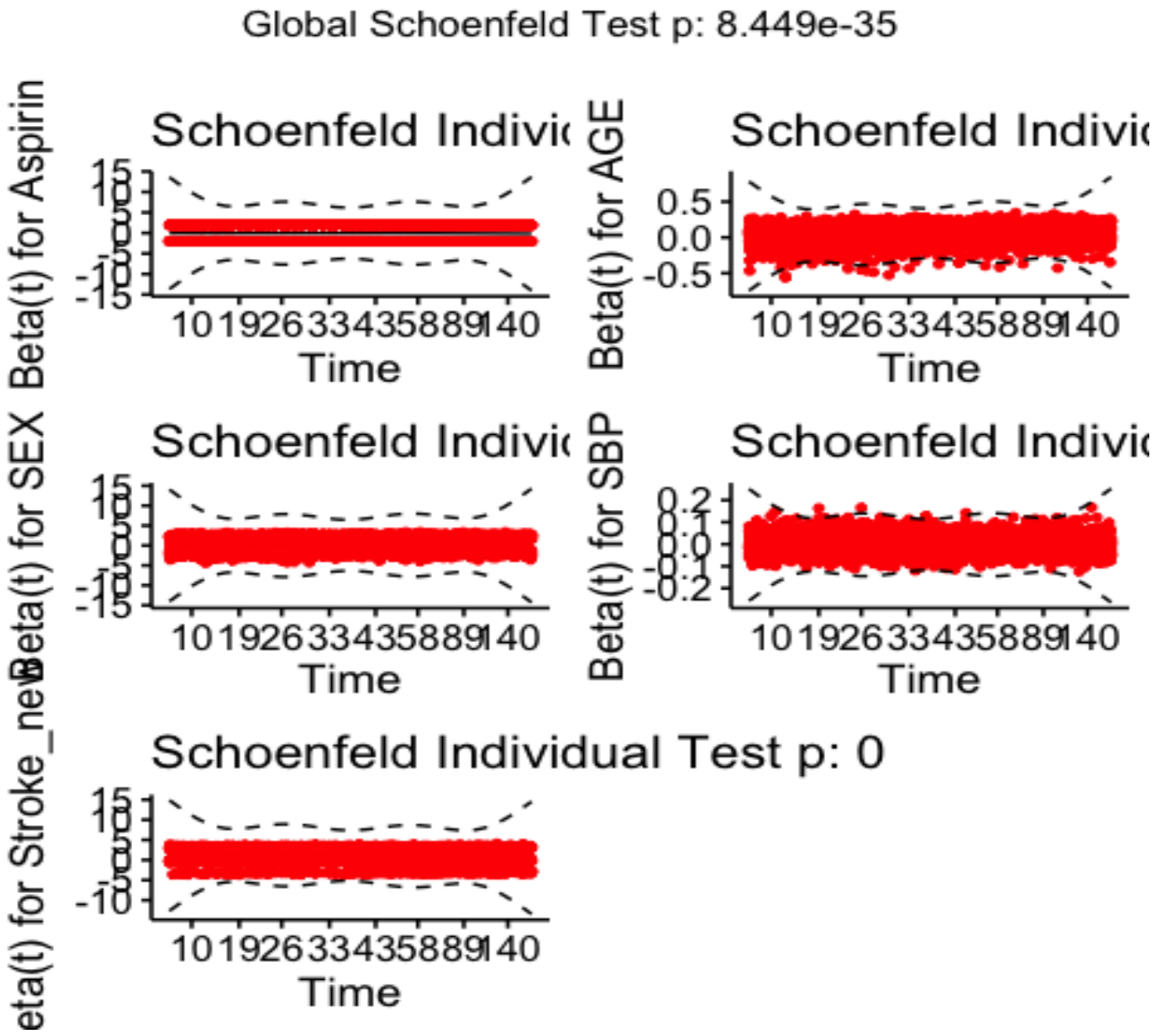
3.5. Assumptions of the Model

For this study, non-informative censoring was assumed, which indicates that patients who are censored at any given time point have the same survival time as those who have similar covariate characteristics but are still at risk. Further, to check the PH assumption Schoenfeld residuals test was employed, the best fitted line for the Schoenfeld residuals for sex, systolic blood pressure, Aspirin Treatment have a slope which is not significantly different from zero with p value = 0.28, 0.82, and 0.31 respectively. Thus, it can be concluded that these covariates do not violate the PH assumption of the fitted model. However, the age and stroke type the slope is significantly different from zero (p-value <0.0001). This indicates that the covariate "age and stroke type" violates the PH assumption of the model and hence leads to the deduction these covariates are probably a time dependent covariate.

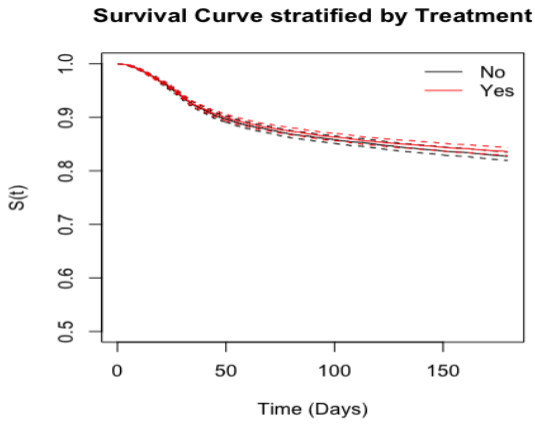
4. Conclusion

It can be argued that aspirin treatment has no substantial effect on patient survival, while factors such as age, stroke type, and systolic blood pressure play a significant role in the hazard. Poor survival is linked with an increase in the age and stroke of the patients. A higher systolic blood pressure value, on the other hand, lowers the risk of mortality.

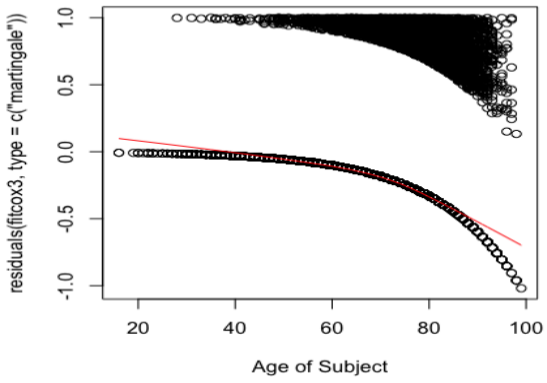
Appendix



Appendix 1 Schoenfeld Test



Appendix 2: Average Survival Curve



Appendix 3: Martingale residual for Age Covariate

Table 2 Variables Selected for the Study

Variables	Description
HOSPNUM	Hospital Number
SEX	M=male; F=Female
AGE	Age in years
RSBP	Systolic blood pressure at randomisation (mmHg)
STYPE	Stroke subtype (TACS/PACS/POCS/LACS/OTH=other)
RXASP	Trial aspirin allocated (Y/N)
Death_Ind_6	Death status at 180 days/6months
Death_c_Ind_6	Indicator for the specific cause of death
STYPE2	Stroke subtype(TACS/PACS/POCS/LACS)
Death_Time_6	Time to Death at 6 Months
Death_C	Death due to Plumunary Embolism and others

Distribution of Task

- Writing of protocol (everyone)
- Coding question 1 to 4 (Emmanuel)
- Coding question 5,6 7 & 9 (Dereje)
- Coding question 8 (model building) everyone.
- Report writing (everyone)

R-Code

```r
#Dereje_emmanuel Survival group project

#PACKAGES
library(car)

library(survival)
library(tidyverse)

library(survminer)

library(caTools)
library(ggplot2)
library(gtsummary)
library(cmprsk)

#Reading and viewing the data
data <- read.table("ITS_Clean.csv", sep = ',', header = TRUE)

head(data)  # six observations

tail(data)  # last six observations

names(data) # names of the variables

dim(data)   # dimensions

str(data)

# selecting variables of interst
select <- c("HOSPNUM","SEX","AGE","RSBP","STYPE","RXASP","Death_Time_6",
          "Death_Ind_6","Death_C_Ind_6", "Death_C", "STYPE2")
data_red <- data[ ,which(names(data) %in% select)]
data_red<- na.omit(data_red)

head(data_red)  # first six observations

names(data_red) # names of the variables of the dataset

dim(data_red)   # dimensions

str(data_red)

#renaming variables
data_red<-data_red %>%rename(Hospital=HOSPNUM ,
                   SBP=RSBP ,Time=Death_Time_6,
                    Status=Death_Ind_6,
                     Death=Death_C_Ind_6,
                      Aspirin =RXASP,
                   Stroke =STYPE,
                   Stroke_new=STYPE2,
```

```
                  Event=Death_C)

#changing the levels of some variables
levels(data_red$Aspirin)[levels(data_red$Aspirin)=="Y"] <- "Yes"
levels(data_red$Aspirin)[levels(data_red$Aspirin)=="N"] <- "No"
levels(data_red$Status)[levels(data_red$Status)=="TRUE"] <- "1"
levels(data_red$Status)[levels(data_red$Status)=="FAlse"] <- "0"

# recode status indicator
data_red$Status<- as.factor(data_red$Status)

#Descriptive analysis for training data
summary(data_red)

table(data_red$Death)

#Histogram of continous variables
hist(data_red$SBP, prob="true", main="Histogram for Systolic blood pressur
e(mmHg) of patients ",
     xlab="SBP",  border="green",  col="blue",
     las=1, breaks=5)

hist(data_red$AGE, main="Histogram for Age of patients ",
     xlab="AGE",  border="green",  col="blue",
     las=1, breaks=5)

hist(data_red$Time, main="Histogram for time to death at 6 month of patien
ts ",
     xlab="Time",  border="green",  col="blue",
     las=1, breaks=5)

#status
data_red$Status<- car::recode(data_red$Status, "'TRUE'=1; 'FALSE'=0;", as.
factor=FALSE)

# overall KM Curve
fit1=survfit(Surv(Time,Status)~1,type="kaplan-meier",conf.type="log-log",
data=data_red)
plot(fit1,  conf.int=TRUE, xlab = "Time (Days) ",ylab = "S",
       main = "Estimated Marginal Survival Curve",ylim = c(0.5,1.0))

#KM stratied by treatment
fit2=survfit(Surv(Time,Status)~Aspirin,type="kaplan-meier",conf.type="log-
log", data=data_red)
ggsurvplot(fit2, data = data_red, conf.int = TRUE, pval = TRUE,
           risk.table = TRUE)

#KM startified by Gender
fit3=survfit(Surv(Time,Status)~SEX,type="kaplan-meier",conf.type="log-log"
,data=data_red)
```

```r
ggsurvplot(fit3, data = data_red, conf.int = TRUE, pval = TRUE,
           risk.table = TRUE)

#Log-rank test by SEX

LR_S=survdiff(formula=Surv(Time,Status)~SEX,
              data=data_red)

#Log-rank test by treatment

LR_T=survdiff(formula=Surv(Time,Status)~Aspirin,
               data=data_red)

#checking the validity of log rank startified by treatment and gender

plot(fit2, xlab="Log(time)", ylab="Log(-log(S(t)))", fun = "cloglog", col=
1:2);
legend("bottomright", names(fit2$strata), bty = "n", col = 1:2,lty = 1)

plot(fit3, xlab="Log(time)", ylab="Log(-log(S(t)))", fun = "cloglog", col=
1:2);
legend("bottomright", names(fit3$strata), bty = "n", col = 1:2,lty = 1)

#Q1
#cumulative hazard

plot(fit2, conf.int=T, fun = function(x) 1-x,
     xlab = "Time in days", ylab = "Cumulative incidence (%)",
     ylim = c(0.0,0.5),col = 2:1,main = "Cumulative Incidence");
legend("topleft", names(fit2$strata), bty = "n", lty = 1, col = 1:2)

#Cox Model to see the effect of treatment
fitcox1<-coxph(Surv(Time, Status) ~ Aspirin, data = data_red)
summary(fitcox1)

coxph(Surv(Time, Status) ~ Aspirin, data = data_red) %>%
        gtsummary::tbl_regression(exp = TRUE)

#Q2
fitcox2<-coxph(Surv(Time, Status) ~ Aspirin+strata(Hospital), data = data_
red)
summary(fitcox2)
str(data_red)

#Q3
fit.ci <- cuminc(data_red$Time, data_red$Event, data_red$Aspirin, cencode=
0)
fit.ci$Tests

ggcompetingrisks(fit.ci)
```

```r
event_c<- data.frame( Event = 0:2,
                          event_char = c("No Event", "Death_C", "Death_Othe
r"))
data_red <- data_red%>%left_join(event_c)

with(data_red, table(Event, event_char))

fit.ci_c <- cuminc(data_red$Time, data_red$event_char, data_red$Aspirin, c
encode="No Event")
ggcompetingrisks(fit.ci_c)

#merging the plot
fit.ci_c2 <- cuminc(data_red$Time, data_red$event_char, cencode="No Event"
)
ggcompetingrisks(fit.ci_c2)

str(data_red)

# cause specific hazard for death due to plumunary disease
death_c <- coxph(formula = Surv(Time, Event==1) ~ Aspirin, data = data_red
)

## cause specific hazard for death due to other cause
death_o <- coxph(formula = Surv(Time, Event==2) ~ Aspirin, data = data_red
)
summary(death_o)

#Q5
fitcox3<-coxph(Surv(Time, Status) ~ AGE, data = data_red) # (significant)
summary(fitcox3)

#5A
#estimated survival curve
survest <-  predict(coxph(Surv(Time, Status) ~ Aspirin + AGE, data=data_re
d), newdata = data_red,type="lp")
median  <-  median(survest)
q1  <-  quantile(survest, 0.25)
q3  <-  quantile(survest, 0.75)
basehaz  <-  basehaz(coxph(Surv(Time, Status) ~ Aspirin + AGE, data=data_r
ed), centered=TRUE)
plot(basehaz$time, exp(-basehaz$hazard*exp(q3)), type='l', xlab = "Prognos
tic score", ylab = "Estimated survival")
lines(basehaz$time, exp(-basehaz$hazard*exp(median)), col='red')
lines(basehaz$time, exp(-basehaz$hazard*exp(q1)), col='blue')
legend("bottomleft", c("Q1", "Median", "Q3"), lty = 1, col = c("blue", "re
d", "black"), bty = "n", cex=1.1)

#probability of surviving for more than 1 year function of age
basehaz12  <-  max(which(basehaz$time<=365))
plot(survest, exp(-basehaz$hazard[basehaz12]*exp(survest)), type='l', xlab
```

```
= "Prognostic score",
     ylab = "Probability of surviving for more than 365 days")



#linearity check
plot(data_red$AGE ,residuals(fitcox3, type=c("martingale")), xlab = "Age o
f Subject")
lines(lowess(data_red$AGE, resid(fitcox3)),col='red')

#Q6 we continue with linear, we added SBP to the linear age

fitcox4<-coxph(Surv(Time, Status) ~ AGE+SBP, data = data_red)
summary(fitcox4)

#Q7

#SBP without age
#Age without SBP

fitcox5<-coxph(Surv(Time, Status) ~ SBP, data = data_red)
fitcox6<-coxph(Surv(Time, Status) ~ AGE, data = data_red)

summary(fitcox5)

summary(fitcox6)

#comparing the two model using AIC
extractAIC(fitcox5)

extractAIC(fitcox6)

#Q8A
fitcox7<-coxph(Surv(Time, Status) ~ Aspirin+AGE+SEX+SBP+Stroke+ Stroke*Asp
irin, data = data_red)

summary(fitcox7)

Anova(fitcox7)

#replacing Stroke with Stroke New
fitcox8<-coxph(Surv(Time, Status) ~ Aspirin+AGE+SEX+SBP+Stroke_new+Stroke_
new*Aspirin, data = data_red)
summary(fitcox8)

#interaction effect (not significant)
Anova(fitcox8)

#Final Model
fitcox9<-coxph(Surv(Time, Status) ~ Aspirin+AGE+SEX+SBP+Stroke_new, data =
data_red)
summary(fitcox9)
```

```r
#table cox model
HR <- round(exp(coef(fitcox9)), 2)
CI <- round(exp(confint(fitcox9)), 2)
SE<-round(coef(summary(fitcox9))[,3], 3)
P <- round(coef(summary(fitcox9))[,5], 3)
# Names the columns of CI
colnames(CI) <- c("Lower", "Higher")
# Bind columns together as dataset
table1 <- as.data.frame(cbind(HR, CI, SE,P))
table1

#assumption PH
final_zph <- cox.zph(fitcox9)
ggcoxzph(cox.zph(fitcox9))

#9
fit10 <- survfit(coxph(Surv(Time, Status)
                       ~ AGE+SEX+SBP+Stroke_new
                       + strata(Aspirin), data=data_red))

plot(fit10, conf.int=T, xlab = "Time (Days)",ylab = "S(t)", main = "Surviv
al Curve stratified by Treatment",
     ylim = c(0.5,1.0),col = c(1:2));
legend("topright", names(fit10$strata), bty = "n", lty = 1, col = c(1:2))
```